動作環境

・Windows 10

・Mac

環境構築

以下の3つが必要です。

・python 3

・gensim(Doc2Vecを使用するためのもの)

・Mecab

1.python 3

pythonは各自環境設定を行ってください。

2.gensim

|  |
| --- |
| pip install gensim |

を実行してください。

(Windows10では、Anaconda Prompt上、Macではターミナル上で実行しました。)

これで、Doc2Vecが使用できるようになります。

3.Mecab

Mecabで形態素解析を行います。

Windows10では以下のサイトを参考に環境設定を行いました。

<https://qiita.com/menon/items/f041b7c46543f38f78f7>

Macでは、brew install mecabでインストールし、mecab-pythonの導入についてはWindowsと同様に設定を行いました。

データセット（教師データ）

githubにソースコードとドキュメントをアップロードしておりますが、テキストデータは著作権の問題上アップロードしておりませんので、各自で用意してください。

今回は教師データとして、以下のURLのwikipediaの「日本史の出来事一覧」のページのテキストをテキストファイルにコピーし使用しました。それを、mecabで形態素解析をして、「名詞」「助動詞」「形容詞」のみをテキストファイルに出力しました。

<https://ja.wikipedia.org/wiki/%E6%97%A5%E6%9C%AC%E5%8F%B2%E3%81%AE%E5%87%BA%E6%9D%A5%E4%BA%8B%E4%B8%80%E8%A6%A7>

センター試験の問題は以下のURL「ロボットは東大に入れるかTodai robot project」からダウンロードしました。各自ダウンロードしてください。

<https://21robot.org/dataset.html>

実行方法

nihonshi.pyとlearn.pyをダウンロードして、pythonで実行してください。

nihonshi.pyは、センター試験の問題文と選択肢を形態素解析した結果をテキストファイルに出力するプログラムです。

learn.pyは、問題文と選択肢の類似度を出力するプログラムです。

オープンしている各テキストファイル名は各自変更してください。

判断方法

実行結果の数値が1に近いほど類似度が高いです。ベクトルが鈍角になっている場合は、値がマイナスになります。